Wide Area Information Servers: A Supercomputer on every Desk

Brewster Kahle
Thinking Machines Corporation

What I really want...

- My personal information to be accessible
- Published information should find me
- Usable anywhere
- Others can use what I have learned (if I want them to)

What is it?

Electronic Publishing

(Or publishing over wires)

New Communications Technology Problems

	BOOKS	Telegraph> Telephone	Electronic Publishing
Experts only	*Monks*	*Operators*	*Professional searchers*
Distribution is hard and expensive	*Vellum is calf skin*	*Telephones on barb wire*	*$1/minute over obscure modems*
Different interfaces	*1000's of languages in Europe alone*	*Switching was manual*	*//query (W5) inform?*
Material is intractable	*Scrolls and manu-scripts were about as random access as musical scores*	*No white pages*	*600 databases on Dialog ~1 Terabyte 140Gbyte at DJ 80GB card catalog at RLG*
Business model needed	*Centralized printing*	*Pay per minute*	*Not understood*

Navigation Techniques: Paper

- Alphabetical Listings (dictionary, Encyclopedia)
- Indices (back of the book and Readers Guide)
- Table of Contents (outlining)
- Citation index
- "Tree of Knowledge"
- Have you read any good books lately?

Navigation Techniques: Computers

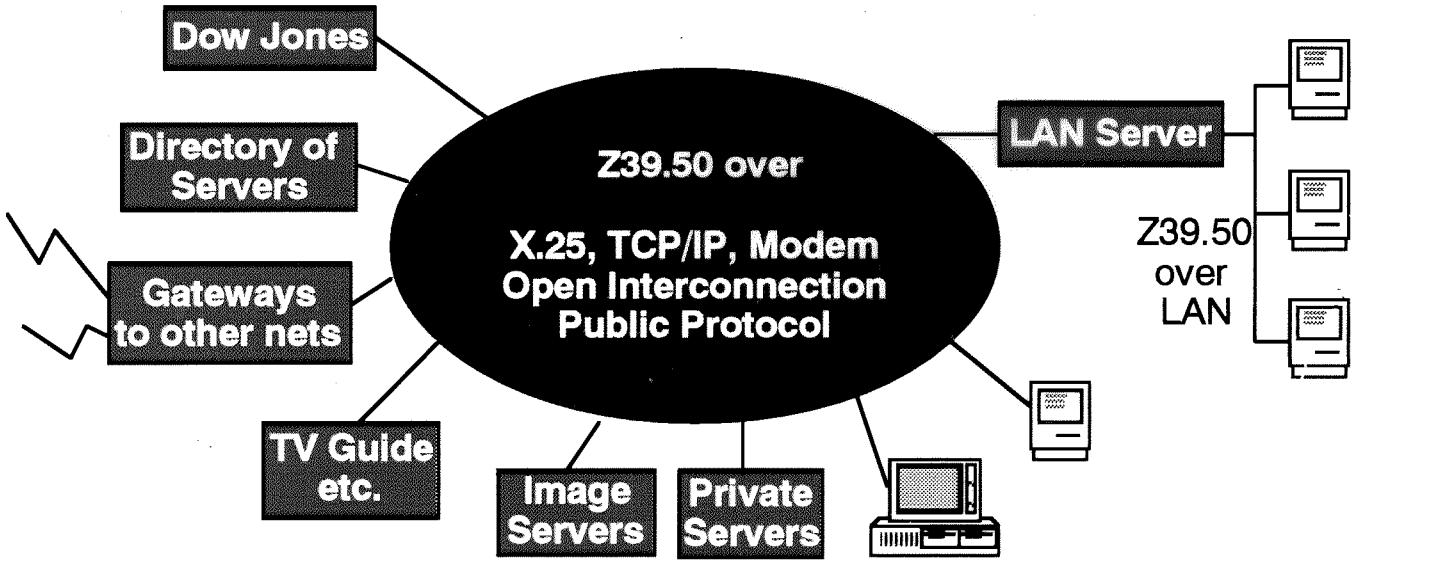
- Hierarchical File Systems
- Unix "find" and "grep", Mac "find file"
- Gopher, Magellan, ON Location
- Boolean query systems (...within 5 words of...)
- Static Hypertext links (see also pointers)

Navigation Techniques: WAIS

- English language questions and relevance feedback
  - Question-answer dialog
  - Similar to Newspapers: "More on page 5"
  - Dynamic Hypertext Links

- 2 level search:
  - Directory of servers (server like any other)
  - Servers themselves

Wide Area Information Server Architecture



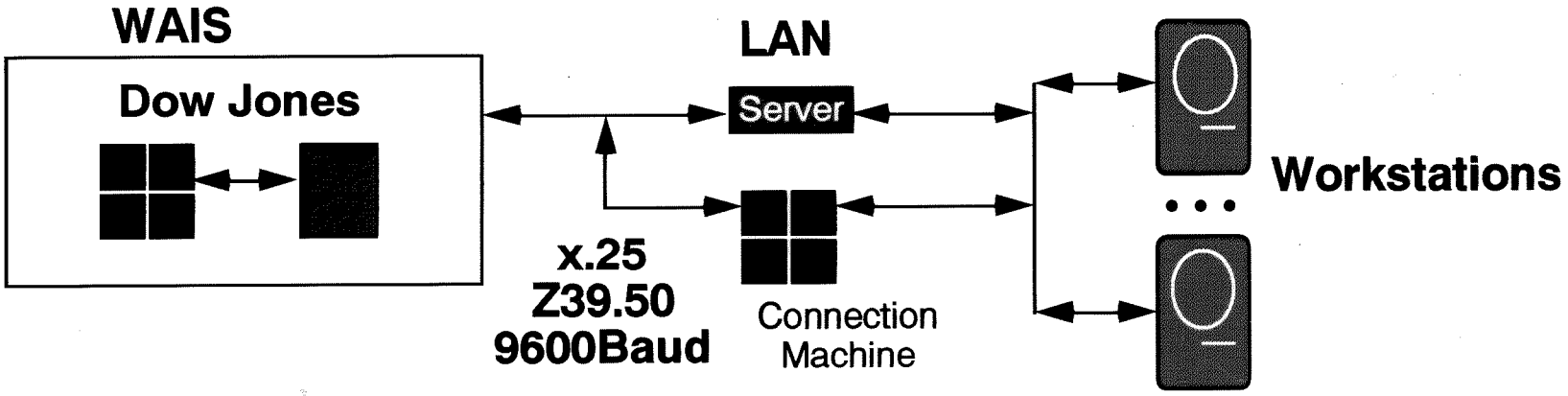
Users Needs:

- Selecting Servers
- Answering Questions
- Organizing Responses

Architecture Issues:

- Scalability
- Security
- Business model for servers
- Reliable Access

Demonstration System Structure



Operations:
- Archiving
- Queries
- Retrieval

IR Type:
- Broadcast
- Query by Example

Databases:
- Wall St Journal
- Barron's
- 400 Business Mags

CM:

Operations:
- Queries

IR Type:
- enhanced relevance feedback

Databases:
- DowVision and memo's, mail, word processor files

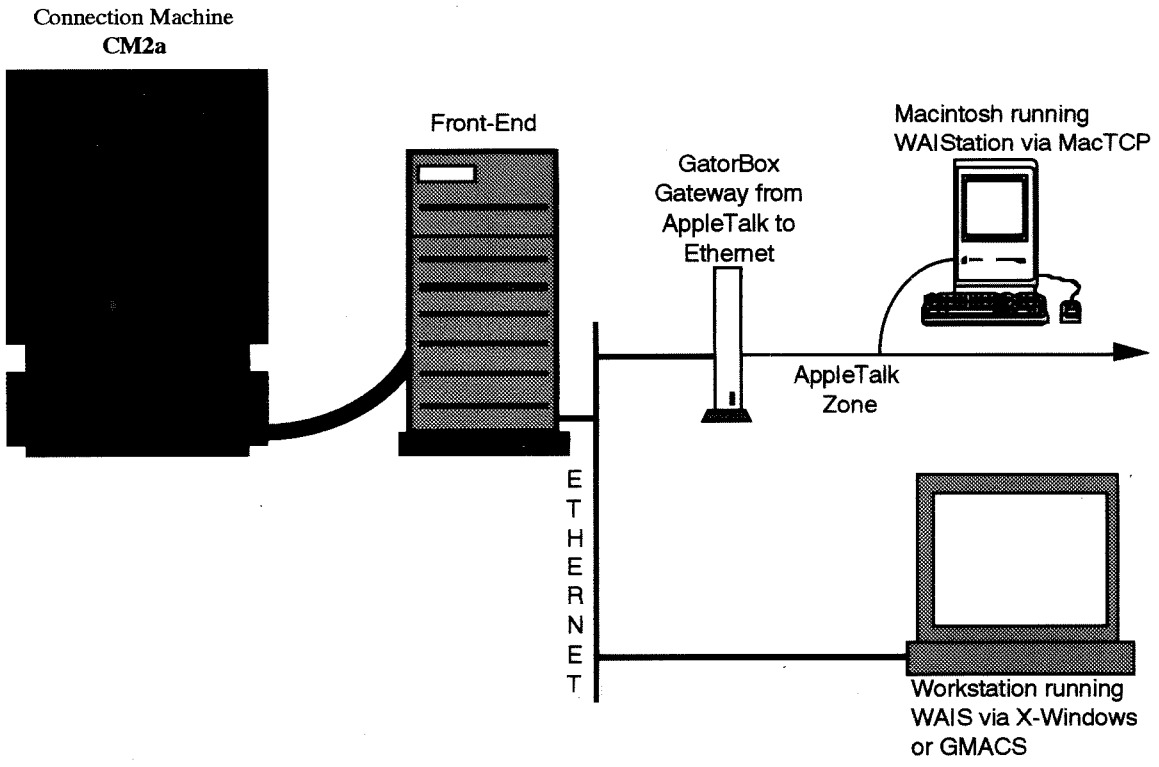
Mac:

Operations:
- Human Int
- Retrieval
- Queries
- "Caching" Docs
- User Profiles

IR Type:
- Query by example

Databases:
- Personal Text
- Cached data

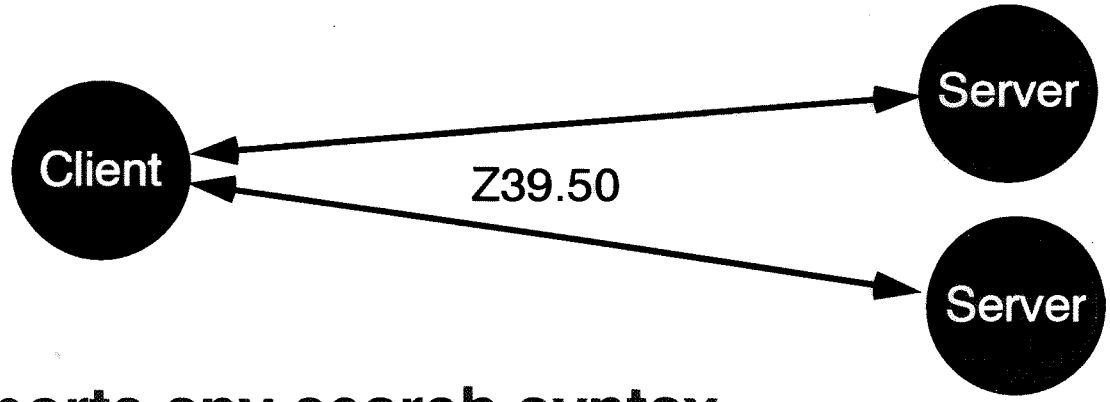
WAIS Hardware Components



WAIS Clients

- Busy 24 hours a day finding information
- Ponder all indications of the preferences of its user
- Gossip with other clients about their discoveries
- Scours the world (within a budget) to find new sources
- Current implementations on PC, Macintosh, X Windows, NeXT, dumb terminal (dial-up)

The WAIS Protocol *is* WAIS



- Supports any search syntax
- Supports sophisticated clients — puts intelligence in the user's hands
- Clients can run on any platform
- Multiple servers in a single search
- Retrieve any kind of data: text, graphics, video,...

WAIS Protocol

- Based on NISO Z39.50 international standard
- Flexible — separates clients from servers
- Search: *(words, doc_ids, databases)*
  returns list of: *(headline, score, doc_id, types)*
- Retrieval: *(doc_id, type, start, end)*
  returns: *data of specified type*
- Doc_id: An ISBN for the Electronic Age
- Server Description Structure for the Directory of Servers

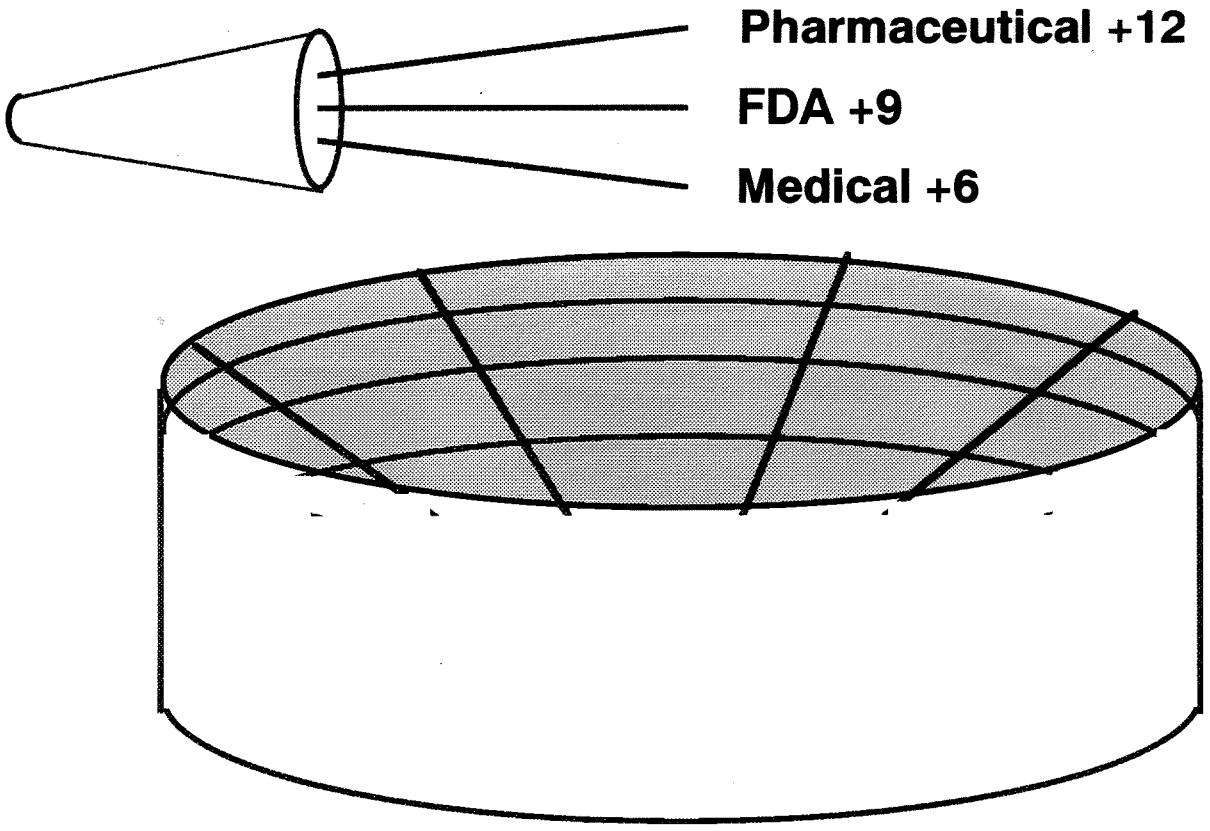
How Standard Protocol can Provides Security

- Users do not login to server, but search only through application layer protocol (Z39.50).
- Server controls access to data.
- Network layers below application, or application layer handles authentication, encryption, billing.

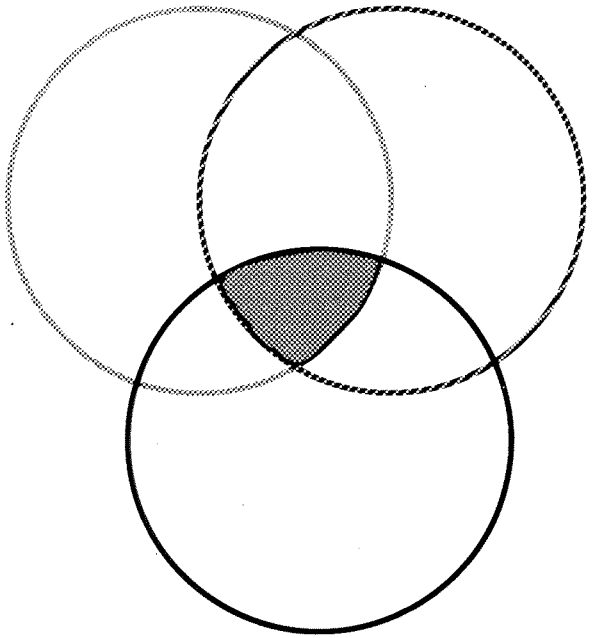
Connection Machine Server

- 1-100GBytes (and getting bigger)
- Supports thousands of users
- Automatic Indexing
- Uses words and phrases in question to find appropriate documents with relevance feedback, weighted term
- Supports Boolean Queries
- Cost effective hardware alternative to mainframes

Data Parallelism:
Searching all the Documents at Once

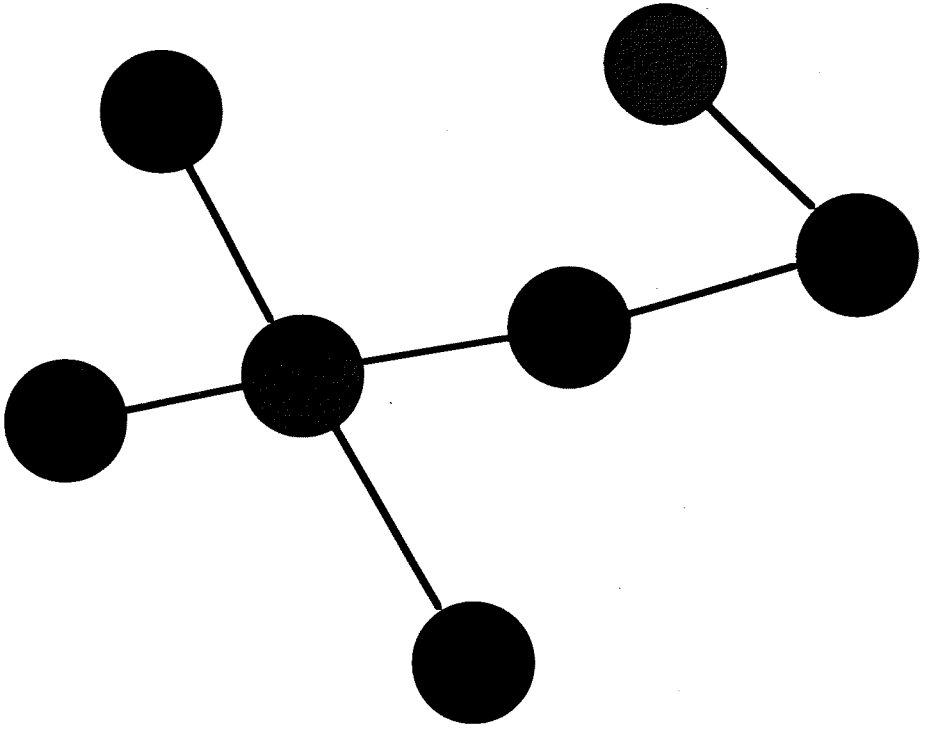


Boolean Search



Retrieve documents containing specific combinations of words

Conceptual Search



Explore a set of documents containing related concepts

Boolean Query

Hard to Use:
Complex Syntax

(Japanese OR Japan) AND
(building OR buildings OR (Real AND Estate) AND
(Manhattan OR (New AND York)

Poor Results:
The wrong information
No ranking of results

Have you been paying attention?...
Freer Finance: U.S. Regulators Move...
REAL ESTATE: California Initiatives...
First Boston Said To Agree on Sale Of...
Exxon, Rockefeller Group to Sell Site...
What's News--Business and Finance

Conceptual Search: Phase 1

Easy to Use:
No Syntax

Japanese buying real estate in mid-town manhattan

Options:
What do you want to follow up?

1. Time Acts to Cut Magazine Costs...
2. *First Boston Said To Agree on Sale...*
3. Have You Been Paying Attention?
4. *Exxon, Rockefeller Group to Sell Site...*
5. Hard Sell: Real Estate Developers...
6. What's News--Business and Finance...
7. Integrated Resources Buys Loft Building...

Conceptual Search: Phase 2

Relevance Feedback:
I like these;
show me more.

First Boston Said To Agree on Sale...
Exxon, Rockefeller Group to Sell Site...

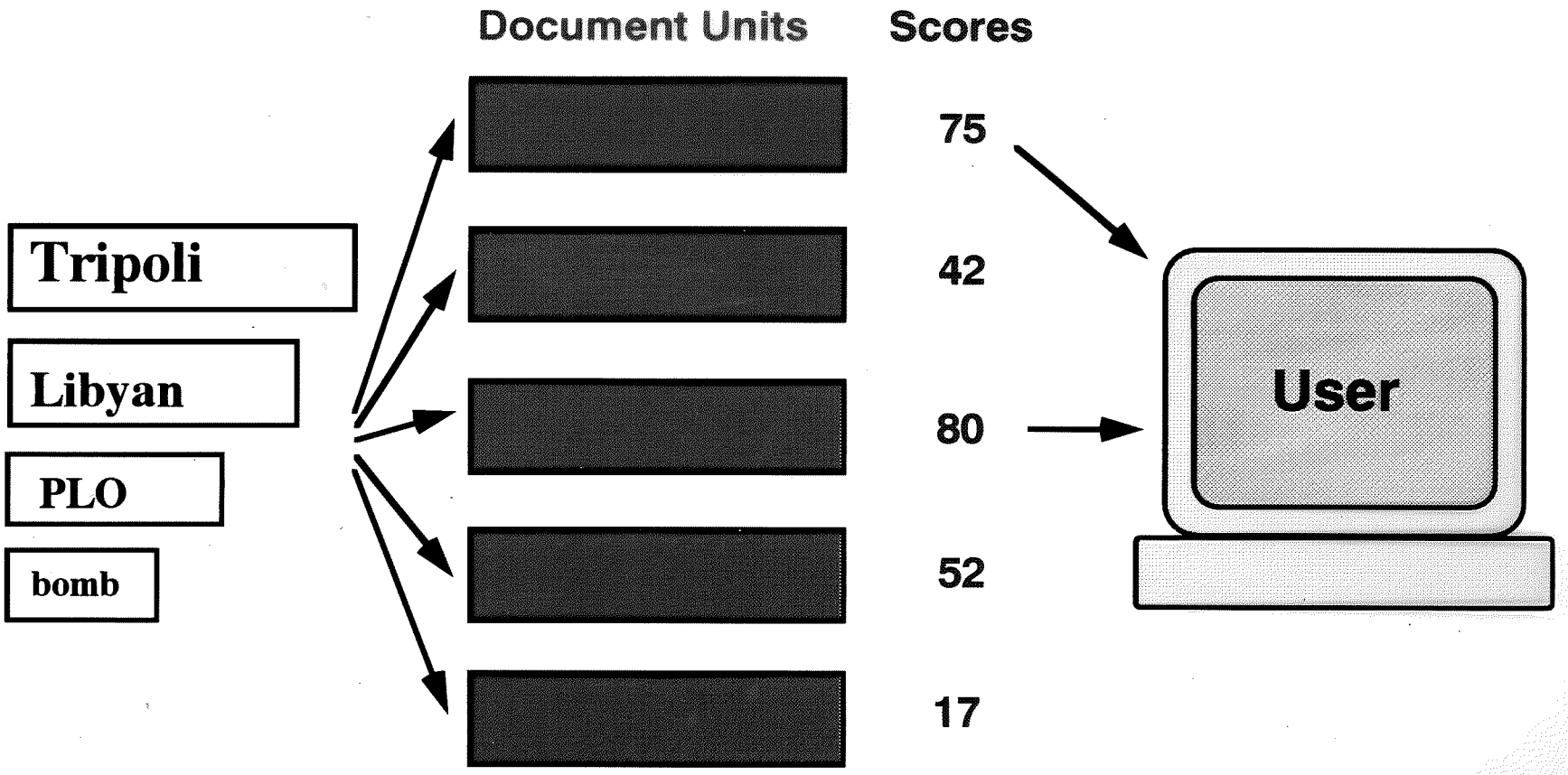
Improved results:

Articles on related topics are found.

Results are ranked.

1. **Bids for Exxon Building in New York...**
2. **Time Acts to Cut Magazine Costs...**
3. **Hard Sell: Real Estate Developers...**
4. **Time Inc. Sells Its 45% Interest...**
5. **Citicorp Unit Moves to Foreclose on...**
6. **Litigious Landlords: Legal Maneuvers**

Query Broadcast To Database on Connection Machine System



Document Retrieval Performance

- Current algorithm limits:
  ~2 GB with 512 MB CM-2
  ~8 GB with 2 GB CM-2
  ~25 GB with 8 GB CM-2

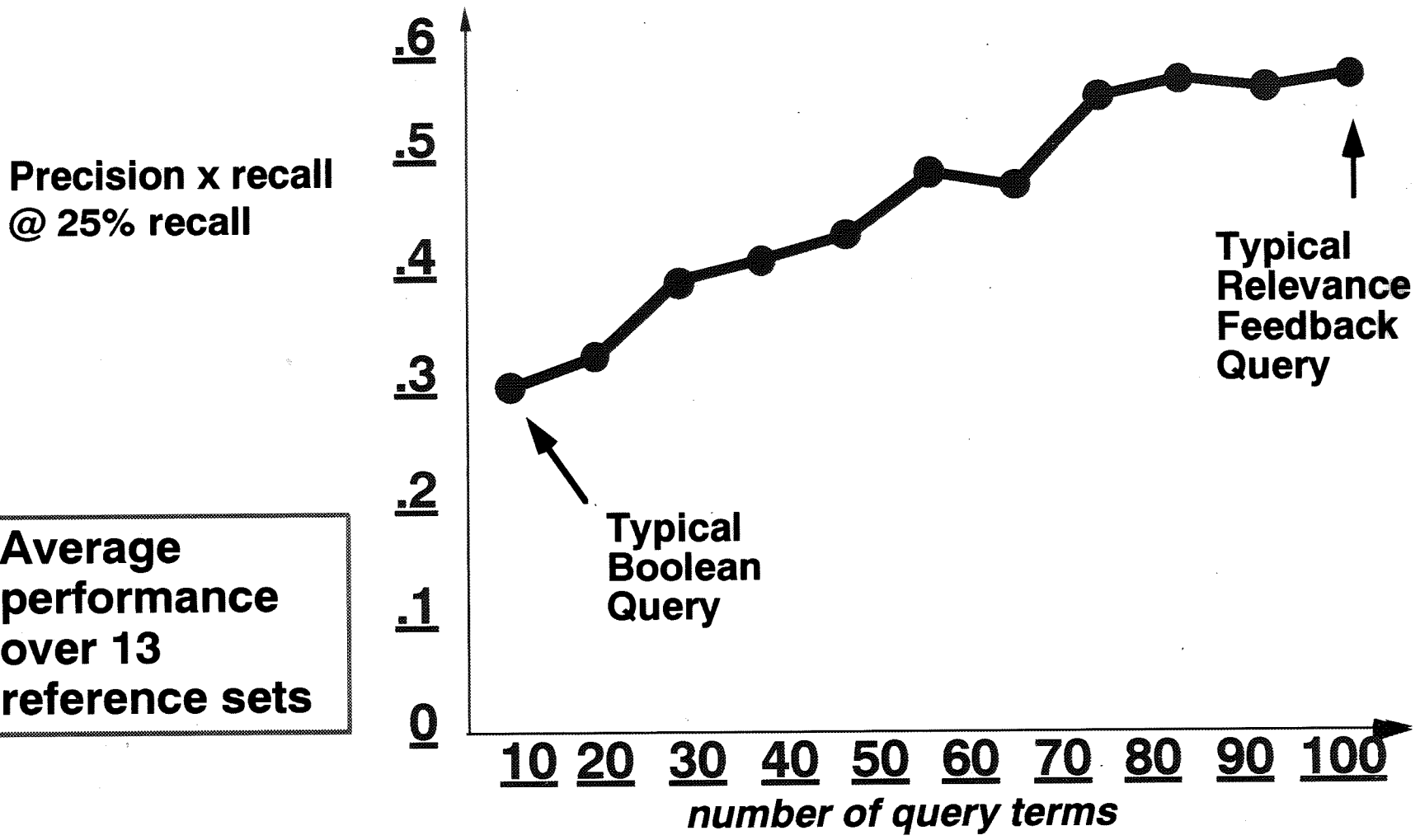
- High recall
- High precision

} see Stanfill and Kahle, *Communications of the ACM*, December 1986

- << 1 sec. response

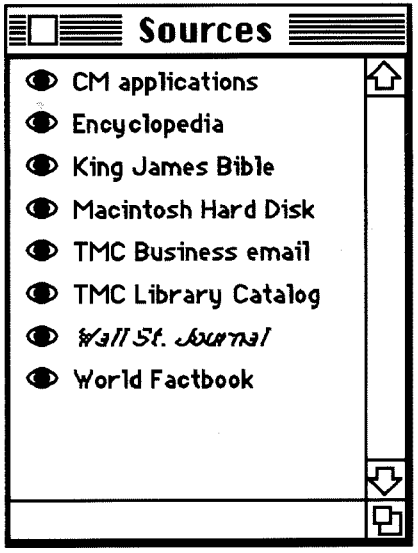
- Much larger DBs searchable with CM-5 and inverted index algorithms: 100s to 1000s of Gigabytes

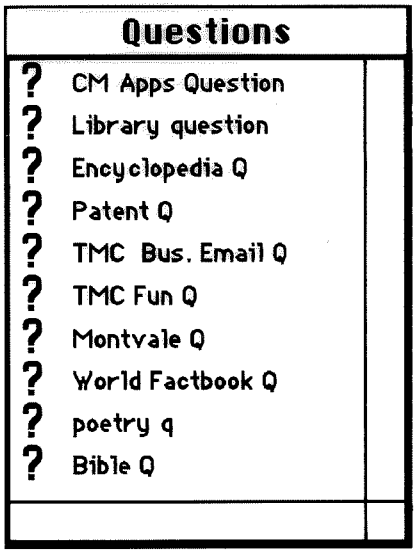
Results Improve with Query Size



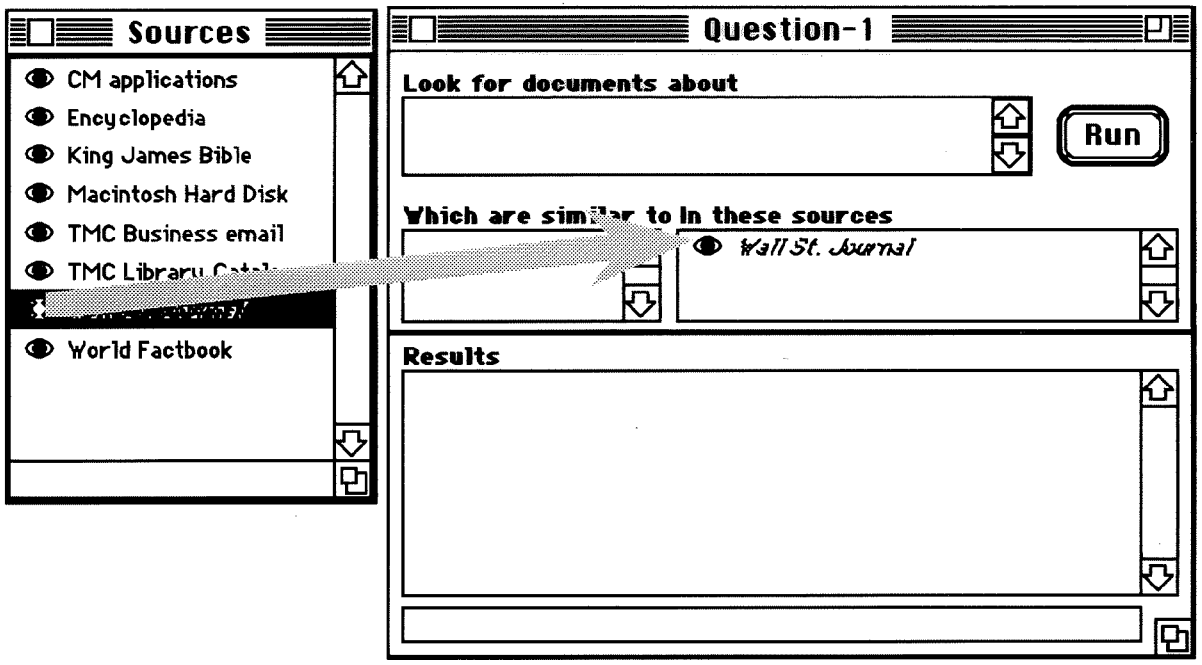
Average performance over 13 reference sets

WAIStation: Active Database Sources, Saved Questions

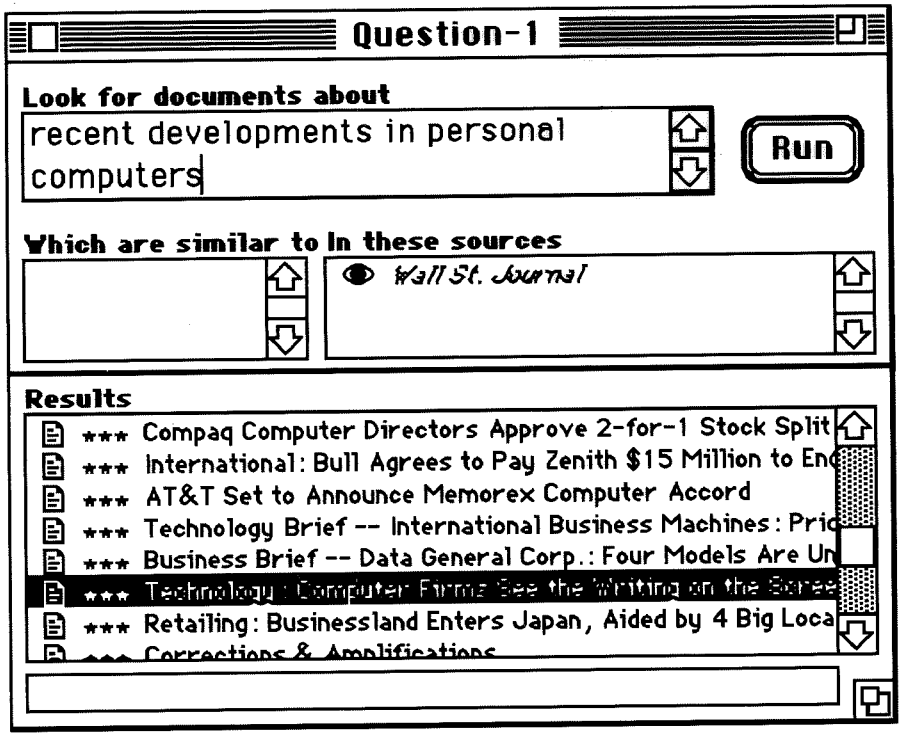




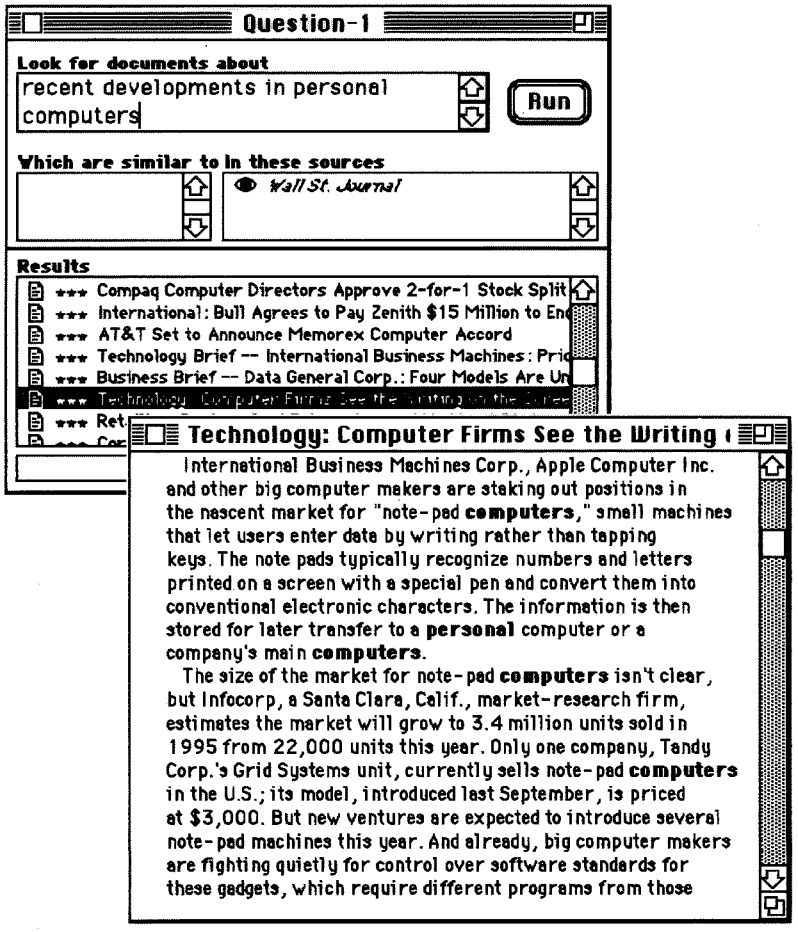
Select Data Source



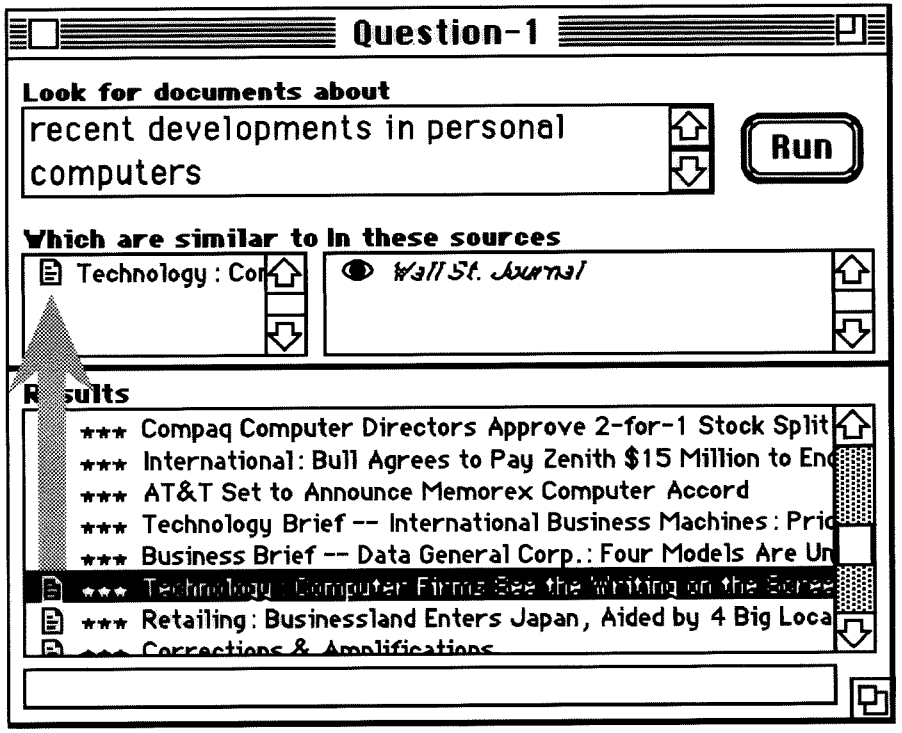
Run Initial Query



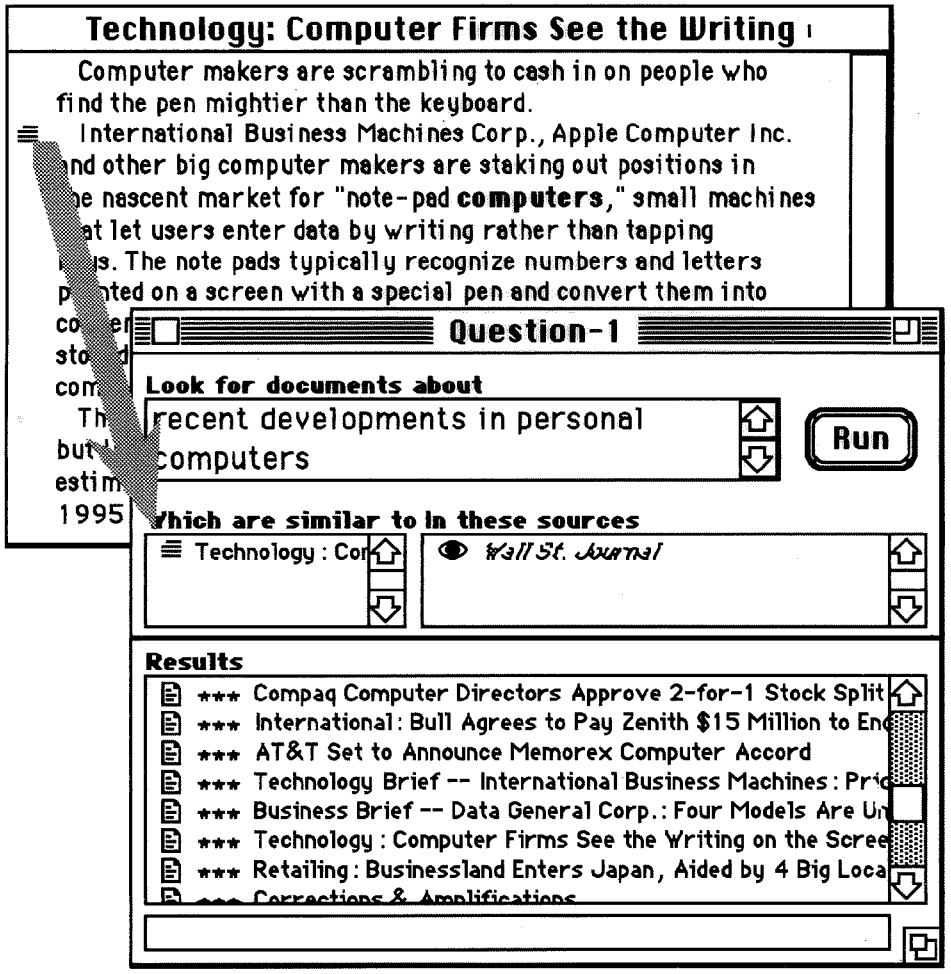
Click a Headline to Display a Document



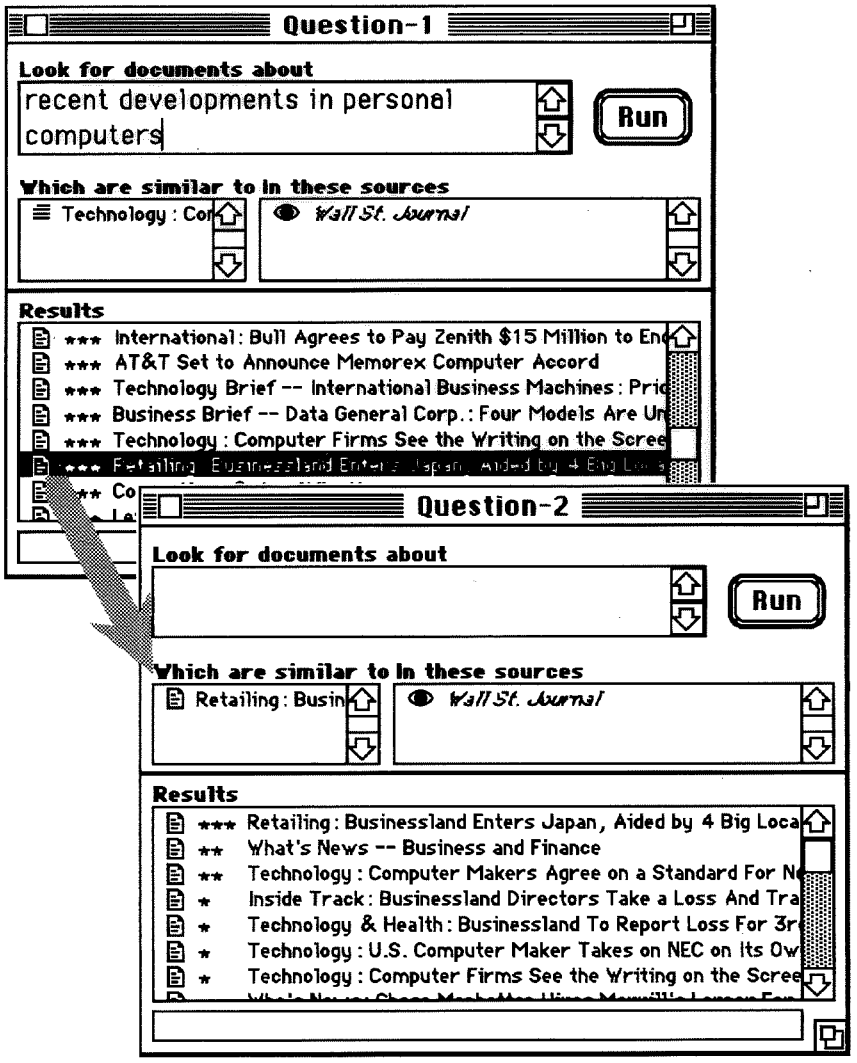
Relevance feedback: "Find me more like this one"



Relevance Feedback of Paragraph



"Chaining" of Questions to Follow a Tangent



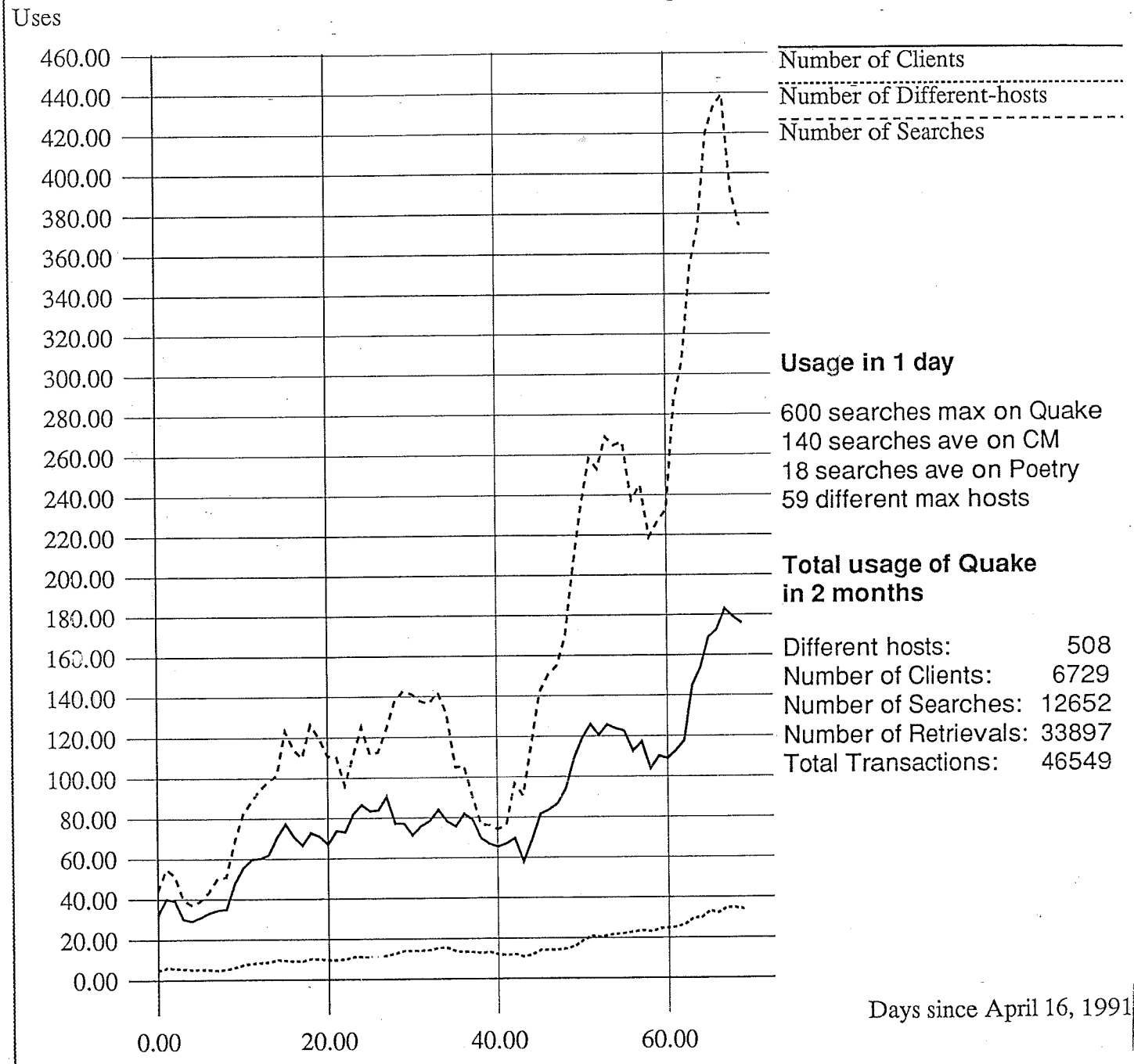
TMC Internet Release

- CM product for TCP/IP (complete server)
- Example User interfaces for free (no support) Macintosh, Gnu Emacs, Xwindows
- Example unix server software to create servers
- Directory of Servers on the internet at least through '92
- 160 Servers now: Weather Maps, patents, journal abstracts, email archives, usenet recipies,...
- Free Software via FTP from Think.com:`/wais/*`

Mailing list: `wais-discussion-request@think.com`

WAIS

WAIS Daily Usages on Quake.Think.Com



Usage in 1 day

600 searches max on Quake
140 searches ave on CM
18 searches ave on Poetry
59 different max hosts

Total usage of Quake in 2 months

Different hosts:	508
Number of Clients:	6729
Number of Searches:	12652
Number of Retrievals:	33897
Total Transactions:	46549

Countries Using WAIS:

Austria, Canada, Denmark, Finland, France, Germany, Holland, Italy, Mexico, Norway, Sweden, Switzerland, USA

Thinking Machines Corporation

WAIS Uses

- Over 10,000 users on the Internet
- Users in 24 Countries: Mexico, Singapore, Finland, Australia, etc
- 160 Databases served from 9 Countries: Norway, Canada, UK, etc. Average 3 new databases registered per week.

WAIS Uses: Libraries

- Easy to use card catalog
- Remote use from home or office
- Pictures, full text, scanned documents

```
[pegun.law.columbia.e]   columbia-law-library-catalog
[pegun.law.columbia.e]   columbia-spanish-law-catalog
[      quake.think.com]   tmc-library
```

WAIS Uses:
Campus Wide Info Servers

- Class catalog and schedule
- Campus events: movies, sports
- Job listings
- Library catalog
- Phone book
- Professor research interests
- Past theses

```
[      sol.acs.unt.edu]  UNTComputerDoc
[        xantos.uio.no]  UiO_Publications
[    next2.oit.unc.edu]  ibm.pc.FAQ
```

WAIS Uses: Biology

- Journal Abstracts
- Sequence archives
- Images

Currently over 20 Biology databases in Finland, Netherlands, and US

```
[       cmns.think.com]  Molecular-biology
[            bio.vu.nl]  biology-compounds
[     genbank.bio.net]   biology-journal-contents
[       wais.funet.fi]   bionic-ai-researchers
[       wais.funet.fi]   bionic-directory-of-servers
[       wais.funet.fi]   bionic-enzyme
```

WAIS Uses: Chemistry CORE Project

- All published chemistry (8 years all ACS)
- Scanned pictures, ascii text
- Optical jukebox mass storage
- Connection Machine / Newton search engines

Project of Bellcore, ACS, Chem Abstracts, OCLC, Cornell, and Thinking Machines

```
[   cujo.curtin.edu.au]   chem-eng-current-contents
```

WAIS Uses: Business Executives

- Dow Jones information
- Corporate information
- Personal information

Project:
KPMG, Apple, Thinking Machines, Dow Jones

```
[       cmns.think.com]  wall-street-journal-sample
[            think.com]  Business-email
```

WAIS Uses: Medical Researchers/Doctors

- Medical papers
- Storing and matching patient records
- Remote connections to specialized databases

```
[        wais.funet.fi]  bionic-databases-limb
```

WAIS Uses: Community Information

- Dial-up users: no network required
- Directories of services or facilities
- Education and entertainment

```
[      quake.think.com]  internet-resource-guide
[      sol.acs.unt.edu]  online-libraries
[      quake.think.com]  weather
[ lambada.oit.unc.edu]   nsf-bulletins
```

Conclusion

- Electronic Publishing can fill niches now.
- Companies are positioning themselves now (workstations, server, and info providers).
- Thinking Machines is the "Engine of the Information Industry."